

Rejection-free Ensemble MCMC with applications to Factorial Hidden Markov Models

Kaspar Märtens *

Michalis K. Titsias [†]

Christopher Yau ^{‡§}

Abstract

Bayesian inference for complex models is challenging due to the need to explore high-dimensional spaces and multimodality and standard Monte Carlo samplers can have difficulties effectively exploring the posterior. We introduce a general purpose rejection-free ensemble Markov Chain Monte Carlo (MCMC) technique to improve on existing poorly mixing samplers. This is achieved by combining parallel tempering and an auxiliary variable move to exchange information between the chains. We demonstrate this ensemble MCMC scheme on Bayesian inference in Factorial Hidden Markov Models. This high-dimensional inference problem is difficult due to the exponentially sized latent variable space. Existing sampling approaches mix slowly and can get trapped in local modes. We show that the performance of these samplers is improved by our rejection-free ensemble technique and that the method is attractive and “easy-to-use” since no parameter tuning is required.

1 Introduction

Monte Carlo-based Bayesian inference (Andrieu et al., 2003) for complex high-dimensional posteriors is a challenging problem as efficient Monte Carlo samplers need to be able to move efficiently across potentially irregular landscapes that may contain many local modes (Frellsen et al., 2016).

In practice, commonly used sampling approaches can explore the space very slowly or get stuck in local modes. For example, Gibbs sampling can have trouble moving around the high-dimensional space, due to the conditional updating scheme where each coordinate is sampled when keeping the rest fixed. In cases where moving between local modes requires a joint update of multiple coordinates, conditional sampling scheme fails to explore the whole space. In a more general Metropolis-Hastings framework, it is difficult to construct efficient proposals for high-dimensional distributions, so the sampling scheme can be inefficient due to the accept-reject process.

We propose a novel ensemble Markov Chain Monte Carlo (MCMC) approach to improve the mixing properties and achieve more effective samplers by using an ensemble of potentially tempered chains, and constructing a rejection-free move to exchange information between them. We demonstrate the benefit of our approach by applying it to challenging inference problems associated with the exponentially sized latent variable spaces of Factorial Hidden Markov Models using a toy sampling problem and a simulation study driven by an application in cancer genomics.

The remainder of the paper is as follows. In Section 2 we discuss ensemble MCMC methods and we introduce our rejection-free scheme. In Section 3 we consider rejection-free ensemble MCMC for inference in Factorial Hidden Markov Models. In Section 4 we demonstrate the practical utility of our methods in a series of numerical experiments and we conclude with a discussion in Section 5.

*Department of Statistics, University of Oxford, United Kingdom

[†]Athens University of Economics and Business, Greece

[‡]Wellcome Trust Centre for Human Genetics, University of Oxford, United Kingdom

[§]Centre for Computational Biology, Institute of Cancer and Genomic Sciences, University of Birmingham, United Kingdom

2 Rejection-free ensemble MCMC

Here, we introduce a general procedure for rejection-free ensemble MCMC (Section 2.2) which extends the standard ensemble methods (Section 2.1).

2.1 Standard ensemble sampling methods

Suppose our goal is to sample from a density π , but our sampler is inefficient and gets stuck in local modes. One solution is to use ensemble MCMC (also known as population-based MCMC, or evolutionary Monte Carlo) methods, which are based on an ensemble of chains (Jasra et al., 2007; Neal, 2011; Shestopaloff & Neal, 2014). That is, a new target density π^* is defined on the product space such that

$$\pi^*(\mathbf{x}_1, \dots, \mathbf{x}_K) = \prod_{k=1}^K \pi_k(\mathbf{x}_k), \quad (1)$$

whereas $\pi_k = \pi$ for at least one index (Jasra et al., 2007). Here we focus on parallel tempering, which introduces a temperature ladder $T_1 < \dots < T_K$ and associates a temperature with each chain. Denoting the inverse temperature $\beta_k = 1/T_k$, we define the tempered targets $\pi_k = \pi^{\beta_k}$. The idea is that high temperature chains explore the space well and do not get stuck, whereas the chain with $T_1 = 1.0$ samples locally precisely from the target. Mostly we would update each chain independently, but occasionally exchange information between the chains.

One approach to exchange information is to propose swapping states between chains of consecutive temperatures and then accept/reject the swap with Metropolis-Hastings (Geyer, 1991; Earl & Deem, 2005). More elaborate approaches can create proposals using genetic algorithms (Liang & Wong, 2000), by proposing crossovers between chains which again requires accepting/rejecting based on the Metropolis-Hastings framework. However, the accept/reject procedure can be inefficient and very sensitive to the choice of the temperature ladder and algorithmic parameter tuning. Thus, it would be highly desirable to combine chains in a rejection-free manner so that any proposal is accepted with probability one. Next, we introduce such a rejection-free sampling move based on auxiliary variables and Gibbs sampling.

2.2 Rejection-free sampling using auxiliary variables

Consider the target in the product space given by eq. (1). Suppose that during MCMC we would like to exchange information between a pair of chains (typically of consecutive temperatures in the ladder) $\pi_i(\mathbf{x}_i)$ and $\pi_j(\mathbf{x}_j)$ where \mathbf{x}_i and \mathbf{x}_j are T -dimensional vectors that indicate the current states of these chains. We introduce two auxiliary variables \mathbf{u} and \mathbf{v} , that live in the same space as \mathbf{x}_i and \mathbf{x}_j , and they are drawn from an auxiliary distribution $p(\mathbf{u}, \mathbf{v} | \mathbf{x}_i, \mathbf{x}_j)$. Without loss of generality we assume that this auxiliary distribution is uniform over all possible one-point crossovers between \mathbf{x}_i and \mathbf{x}_j . More precisely, given two vectors $x_{1:T}$ and $y_{1:T}$ we define their one-point crossover at point t , where $t = 1, \dots, T$, as follows:

$$\begin{pmatrix} x_1, \dots, x_t, x_{t+1}, \dots, x_T \\ y_1, \dots, y_t, y_{t+1}, \dots, y_T \end{pmatrix} \implies \begin{pmatrix} y_1, \dots, y_t, x_{t+1}, \dots, x_T \\ x_1, \dots, x_t, y_{t+1}, \dots, y_T \end{pmatrix}$$

We also introduce the set $\text{CR}(\mathbf{x}, \mathbf{y})$ to denote all T crossovers between the vectors \mathbf{x} and \mathbf{y} . The auxiliary distribution $p(\mathbf{u}, \mathbf{v} | \mathbf{x}_i, \mathbf{x}_j)$ is precisely an uniform distribution over all pairs $(\mathbf{u}, \mathbf{v}) \in \text{CR}(\mathbf{x}_i, \mathbf{x}_j)$. This distribution is also symmetric, i.e.

$$p(\mathbf{u}, \mathbf{v} | \mathbf{x}_i, \mathbf{x}_j) = p(\mathbf{x}_i, \mathbf{x}_j | \mathbf{u}, \mathbf{v}),$$

which means that when we condition on fixed values (\mathbf{x}, \mathbf{y}) the above is an uniform distribution over all pairs $(\mathbf{u}, \mathbf{v}) \in \text{CR}(\mathbf{x}_i, \mathbf{x}_j)$, while when we condition on fixed (\mathbf{u}, \mathbf{v}) the above reduces to an uniform distribution over all pairs $(\mathbf{x}_i, \mathbf{x}_j) \in \text{CR}(\mathbf{u}, \mathbf{v})$.

Using the auxiliary variables we can exchange information between \mathbf{x}_i and \mathbf{x}_j through the intermediate step of sampling the auxiliary variables (\mathbf{u}, \mathbf{v}) based on the following two-step Gibbs procedure:

1. Generate $(\mathbf{u}, \mathbf{v}) \sim p(\mathbf{u}, \mathbf{v} | \mathbf{x}_i, \mathbf{x}_j)$
2. Generate $(\mathbf{x}_i, \mathbf{x}_j) \sim p(\mathbf{x}_i, \mathbf{x}_j | \text{rest})$ where

$$\begin{aligned} p(\mathbf{x}_i, \mathbf{x}_j | \text{rest}) &= \frac{1}{Z} \pi_i(\mathbf{x}_i) \pi_j(\mathbf{x}_j) p(\mathbf{u}, \mathbf{v} | \mathbf{x}_i, \mathbf{x}_j) \\ &= \frac{1}{Z} \pi_i(\mathbf{x}_i) \pi_j(\mathbf{x}_j) p(\mathbf{x}_i, \mathbf{x}_j | \mathbf{u}, \mathbf{v}) \\ &= \frac{1}{Z} \pi_i(\mathbf{x}_i) \pi_j(\mathbf{x}_j) I((\mathbf{x}_i, \mathbf{x}_j) \in \text{CR}(\mathbf{u}, \mathbf{v})) \end{aligned}$$

where the normalising constant $Z = Z(\mathbf{u}, \mathbf{v})$ is

$$Z(\mathbf{u}, \mathbf{v}) = \sum_{(\mathbf{y}_i, \mathbf{y}_j) \in \text{CR}(\mathbf{u}, \mathbf{v})} \pi_i(\mathbf{y}_i) \pi_j(\mathbf{y}_j).$$

The first step of the above procedure selects a random crossovered pair (\mathbf{u}, \mathbf{v}) , while the second step conditions on this selected pair and jointly samples $(\mathbf{x}_i, \mathbf{x}_j)$ from the exact conditional posterior distribution that takes into account the information coming from the actual chains π_i and π_j .

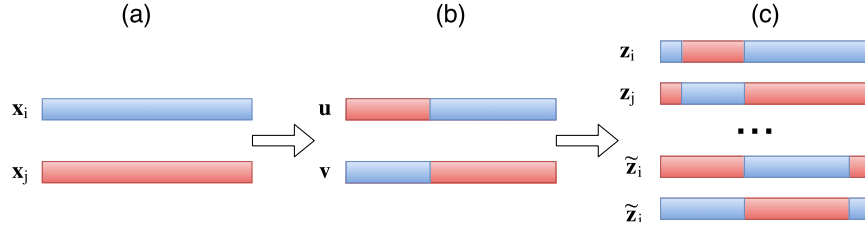


Figure 1: Schematic overview of the auxiliary variable crossover move. (a) We start with two sequences \mathbf{x}_i and \mathbf{x}_j . (b) Now we construct auxiliary variables \mathbf{u}, \mathbf{v} by applying a uniform one-point crossover to $\mathbf{x}_i, \mathbf{x}_j$. (c) Next, we consider all possible crossovers of \mathbf{u}, \mathbf{v} and calculate posterior probabilities $\pi_i(\mathbf{z}_i) \pi_j(\mathbf{z}_j)$ for all of them. Finally, we accept one of these configurations as the new value of $\mathbf{x}_i, \mathbf{x}_j$.

Since the above is a Gibbs operation it is quite obvious that it should lead to new state vectors for the chains π_i and π_j that are always accepted. To prove this more explicitly next we compute the effective marginal proposal, by marginalising out the auxiliary variables, and show that the corresponding Metropolis-Hastings acceptance probability is always one.

Given the current states $(\mathbf{x}_i, \mathbf{x}_j)$, we denote the proposed states by $(\mathbf{z}_i, \mathbf{z}_j)$ and the marginal proposal distribution by $Q(\mathbf{z}_i, \mathbf{z}_j | \mathbf{x}_i, \mathbf{x}_j)$. This proposal, defined by the above two-step Gibbs procedure, is a mixture:

$$\begin{aligned} Q(\mathbf{z}_i, \mathbf{z}_j | \mathbf{x}_i, \mathbf{x}_j) &= \\ &= \iint \frac{1}{Z} \pi_i(\mathbf{x}_i) \pi_j(\mathbf{x}_j) p(\mathbf{z}_i, \mathbf{z}_j | \mathbf{u}, \mathbf{v}) p(\mathbf{u}, \mathbf{v} | \mathbf{x}_i, \mathbf{x}_j) d\mathbf{u} d\mathbf{v} \\ &= \pi_i(\mathbf{x}_i) \pi_j(\mathbf{x}_j) \iint \frac{1}{Z} p(\mathbf{z}_i, \mathbf{z}_j | \mathbf{u}, \mathbf{v}) p(\mathbf{x}_i, \mathbf{x}_j | \mathbf{u}, \mathbf{v}) d\mathbf{u} d\mathbf{v} \\ &= \pi_i(\mathbf{x}_i) \pi_j(\mathbf{x}_j) H(\mathbf{z}_i, \mathbf{z}_j | \mathbf{x}_i, \mathbf{x}_j). \end{aligned}$$

The Metropolis-Hastings acceptance probability under this proposal is

$$\begin{aligned}\alpha &= \frac{p(\mathbf{z}_i, \mathbf{z}_j)Q(\mathbf{x}_i, \mathbf{x}_j|\mathbf{z}_i, \mathbf{z}_j)}{p(\mathbf{x}_i, \mathbf{x}_j)Q(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j)} \\ &= \frac{\pi_i(\mathbf{z}_i)\pi_j(\mathbf{z}_j)Q(\mathbf{x}_i, \mathbf{x}_j|\mathbf{z}_i, \mathbf{z}_j)}{\pi_i(\mathbf{x}_i)\pi_j(\mathbf{x}_j)Q(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j)},\end{aligned}$$

which, since $H(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j) = H(\mathbf{x}_i, \mathbf{x}_j|\mathbf{z}_i, \mathbf{z}_j)$ due to symmetry, simplifies to

$$\alpha = \frac{\pi_i(\mathbf{z}_i)\pi_j(\mathbf{z}_j)\pi_i(\mathbf{x}_i)\pi_j(\mathbf{x}_j)H(\cdot, \cdot|\cdot, \cdot)}{\pi_i(\mathbf{x}_i)\pi_j(\mathbf{x}_j)\pi_i(\mathbf{z}_i)\pi_j(\mathbf{z}_j)H(\cdot, \cdot|\cdot, \cdot)},$$

where all terms cancel out. So $\alpha = 1$ and our proposal will be always accepted.

To simulate from $Q(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j)$ in practice, we can use its mixture representation above, i.e. first generate auxiliary variables $p(\mathbf{u}, \mathbf{v}|\mathbf{x}_i, \mathbf{x}_j)$ and then conditional on those, generate the new value from $p(\mathbf{z}_i, \mathbf{z}_j|\mathbf{u}, \mathbf{v})$. We note that even though both of these steps are implemented as one-point crossovers, the overall proposal can lead to a two-point crossover as illustrated in Figure 1.

A further extension of the above procedure is obtained by modifying the auxiliary distribution $p(\mathbf{u}, \mathbf{v}|\mathbf{x}_i, \mathbf{x}_j)$ to become uniform over the union of the sets $\text{CR}(\mathbf{x}_i, \mathbf{x}_j)$ and $\text{CR}(\mathbf{x}_j, \mathbf{x}_i)$ since, due to the deterministic ordering, the crossovers between \mathbf{x}_i with \mathbf{x}_j and the reverse crossovers between \mathbf{x}_j with \mathbf{x}_i are not the identical. The auxiliary distribution $p(\mathbf{u}, \mathbf{v}|\mathbf{x}_i, \mathbf{x}_j)$ still remains symmetric and all above properties hold unchanged. The only difference is that now we are considering $2T$ crossovers and in order to sample from $p(\mathbf{u}, \mathbf{v}|\mathbf{x}_i, \mathbf{x}_j)$ we need first to flip a coin to decide the order of \mathbf{x}_i and \mathbf{x}_j . Equivalently, we can sample from the initial auxiliary distribution and then randomly assign the outcome to either (\mathbf{u}, \mathbf{v}) or (\mathbf{v}, \mathbf{u}) . Complete pseudocode of the whole procedure is given by Algorithm 1.

Algorithm 1 Scheme for a rejection-free two-point crossover between \mathbf{x}_i and \mathbf{x}_j

```
Pick  $t$  uniformly  $t \sim U(\{1, \dots, T\})$ 
# Flip a coin to decide the direction of crossover
if  $u < 0.5$  where  $u \sim U(0, 1)$  then
     $(\mathbf{u}, \mathbf{v}) \leftarrow \text{CROSSOVER}(\mathbf{x}_i, \mathbf{x}_j, t)$ 
else
     $(\mathbf{v}, \mathbf{u}) \leftarrow \text{CROSSOVER}(\mathbf{x}_i, \mathbf{x}_j, t)$ 
end if
# consider all normal and flipped crossovers of  $u$  and  $v$ 
for  $t \in \{1, \dots, T\}$  do
    # Normal crossover of  $u$  and  $v$ 
     $(\mathbf{z}_i, \mathbf{z}_j) \leftarrow \text{CROSSOVER}(\mathbf{u}, \mathbf{v}, t)$ 
     $a_t \leftarrow \pi_i(\mathbf{z}_j)\pi_j(\mathbf{z}_i)$ 
    # Flipped crossover of  $u$  and  $v$ 
     $(\mathbf{z}_j, \mathbf{z}_i) \leftarrow \text{CROSSOVER}(\mathbf{u}, \mathbf{v}, t)$ 
     $a_{T+t} \leftarrow \pi_i(\mathbf{z}_j)\pi_j(\mathbf{z}_i)$ 
end for
 $a_t \leftarrow a_t / \sum_s a_s$  # Normalise the probabilities
# Pick index  $t_0$  with probability proportional to  $a_{t_0}$ 
 $t_0 \sim \text{Discrete}(a_1, \dots, a_T, a_{T+1}, \dots, a_{2T})$ 
if  $t_0 \leq T$  then
     $(\mathbf{x}_i, \mathbf{x}_j) \leftarrow \text{CROSSOVER}(\mathbf{x}_i, \mathbf{x}_j, t_0)$ 
else
     $(\mathbf{x}_j, \mathbf{x}_i) \leftarrow \text{CROSSOVER}(\mathbf{x}_i, \mathbf{x}_j, t_0)$ 
end if
```

Algorithm 2 Function for a one-point crossover at point t

```
function CROSSOVER( $x_{1:T}, y_{1:T}, t$ )  
   $u_{1:T} \leftarrow (y_1, \dots, y_t, x_{t+1}, \dots, x_T)$   
   $v_{1:T} \leftarrow (x_1, \dots, x_t, y_{t+1}, \dots, y_T)$   
  return( $u_{1:T}, v_{1:T}$ )  
end function
```

The above sampling scheme is very general and it can be applied to arbitrary MCMC inference problems involving both continuous and discrete variables. In the next section we apply the proposed method to a challenging inference problem in Factorial Hidden Markov Models (FHMMs).

3 Application to Factorial Hidden Markov Models

We start with a brief overview of basic Hidden Markov Models (HMMs) and FHMMs in Section 3.1, then we discuss the current state-of-the-art sampling schemes for FHMMs in Section 3.2 and then we apply the rejection-free method to FHMMs in Section 3.3.

3.1 HMMs and FHMMs

HMMs are widely and successfully used for modelling sequential data across a range of areas, including signal processing (Crouse et al., 1998), genetics and computational biology (Marchini & Howie, 2010; Yau, 2013). The HMM assumes that there is an underlying unobserved Markov chain with a finite number of states, which generates a sequence of observations $y_{1:T} := (y_1, \dots, y_T)$ via a parametric emission distribution.

Inference over the latent sequence $x_{1:T}$ and the parameters can be carried out either from a likelihood (Rabiner & Juang, 1986) or Bayesian (Scott, 2002) perspective. The latter is desirable due to the interest in capturing uncertainty over the hidden sequences and is the premise for this paper. Standard inference approaches employ conditional sampling where the parameters and latent sequences are updated iteratively conditional on the other being fixed and latent sequences can be jointly sampled using the forward-filtering-backward-sampling (FF-BS) (Scott, 2002).

The Factorial HMM (FHMM) is an extended version of the standard HMM where instead of a single latent chain, there are K latent chains. That is, given observations $y_{1:T}$, our goal is to infer a $K \times T$ latent matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$ whose columns evolve according to Markov transitions. Here we focus on the case where \mathbf{X} is binary, in which case the element $x_{k,t}$ indicates whether latent feature k contributes to observation y_t or not.

Each observation y_t is generated conditional on the \mathbf{x}_t value from an emission density $p(y_t|\mathbf{x}_t)$ with some parameters ϕ . We treat the rows of \mathbf{X} independently, so denoting the transition probability for chain k by ρ_k , i.e.

$$p(x_{k,t}|x_{k,t-1}) = \begin{cases} 1 - \rho_k & \text{if } x_{k,t} = x_{k,t-1} \\ \rho_k & \text{if } x_{k,t} \neq x_{k,t-1} \end{cases}$$

we can express $p(\mathbf{X})$ as follows

$$p(\mathbf{X}) = \prod_{k=1}^K p(x_{k,1}) \prod_{t=2}^T p(x_{k,t}|x_{k,t-1}).$$

The joint distribution of $(\mathbf{X}, y_{1:T})$ is given by

$$p(y_{1:T}, \mathbf{X}) = \left(\prod_{t=1}^T p(y_t|\mathbf{x}_t) \right) p(\mathbf{X})$$

Even though FF-BS is an exact sampling algorithm, in practice this is infeasible for FHMMs, where the state space grows exponentially in the number of latent sequences K . This results in the complexity of FF-BS being $O(2^{2K}T)$, so even for relatively small values of K there is a need for an alternative approach. Variational inference (Ghahramani et al., 1997) offers a faster alternative for approximate inference but also suffers from scalability issues as well as the problems with multimodality in the posterior distribution.

3.2 Existing MCMC approaches for inference in FHMMs

The standard FF-BS step has quadratic complexity w.r.t the size of the state space, so full FF-BS is not feasible for FHMMs. Thus a computationally cheaper approach is needed, however this comes at the expense of sampling efficiency.

One option is to sample each row of \mathbf{X} conditional on the rest, using the FF-BS. Then each of the updates has a state space of size 2 and the FF-BS steps are inexpensive. However, in this conditional scheme most of the sequences are fixed and thus it is difficult for the sampler to explore the space well. A more general version of this would update a small subset of chains at a time at a higher computational cost, but it can still get trapped in local modes.

An alternative idea referred to as Hamming Ball sampling has been suggested by Titsias and Yau 2014; 2016, which adaptively truncates the space via an auxiliary variable scheme. Unlike the conditional Gibbs updates, it does not restrict parts of \mathbf{X} to be fixed during sampling. Even though it can be less prone to get stuck, for a moderate value of K it may still not explore the whole posterior space.

3.3 Ensemble MCMC applied to FHMMs

Here, we apply the ensemble crossover scheme to FHMMs in order to significantly improve poorly mixing samplers. We achieve this via an ensemble of chains over suitably defined tempered posteriors.

For a latent variable model, one can either temper the whole joint distribution or just the emission likelihood. We chose the latter, so the target posterior of interest becomes

$$\pi_k(\mathbf{X}) := p(\mathbf{X}|y_{1:T}) \propto p(\mathbf{X})p(y_{1:T}|\mathbf{X})^{\beta_k}$$

where \mathbf{X} is a $K \times T$ binary matrix.

As the ensemble crossover scheme was defined on vectors, there are multiple ways to extend this to matrices. Here, we consider a crossover move on the *rows* of the matrix \mathbf{X} , as illustrated in Figure 2 for a one-row crossover.

The core computational step of the algorithm is to compute quantities a_t for all crossover points t . We show that these can be computed recursively in an efficient way. Let $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(j)}$ be the current states for chains i and j , and let \mathbf{U} and \mathbf{V} be the respective matrices after the crossover move. Comparing their crossovers at two consecutive points $t-1$ and t , denoted by $(\mathbf{Z}_{t-1}^{(i)}, \mathbf{Z}_{t-1}^{(j)})$ and $(\mathbf{Z}_t^{(i)}, \mathbf{Z}_t^{(j)})$, we note that these can differ just in column t :

$$\begin{aligned}\mathbf{Z}_{t-1}^{(i)} &:= (\mathbf{v}_1, \dots, \mathbf{v}_{t-1}, \mathbf{u}_t, \mathbf{u}_{t+1}, \dots, \mathbf{u}_T) \\ \mathbf{Z}_t^{(i)} &:= (\mathbf{v}_1, \dots, \mathbf{v}_{t-1}, \mathbf{v}_t, \mathbf{u}_{t+1}, \dots, \mathbf{u}_T)\end{aligned}$$

As a result, the values $a_t = \pi_i(\mathbf{Z}_t^{(i)})\pi_j(\mathbf{Z}_t^{(j)})$ can be computed recursively. Indeed, given the previous value of $\pi_i(\mathbf{Z}_{t-1}^{(i)})$, we can compute $\pi_i(\mathbf{Z}_t^{(i)})$ by accounting for the following two cases:

- the change in emission likelihood from $p(y_t|\mathbf{u}_t)^{\beta_i}$ to $p(y_t|\mathbf{v}_t)^{\beta_i}$
- the change in transitions from $\mathbf{v}_{t-1} \rightarrow \mathbf{u}_t \rightarrow \mathbf{u}_{t+1}$ to $\mathbf{v}_{t-1} \rightarrow \mathbf{v}_t \rightarrow \mathbf{u}_{t+1}$

		$\mathbf{X}^{(i)}$								$\mathbf{X}^{(j)}$					
		1	1	1	1	1	1			0	0	0	0	1	1
(a)		1	1	1	1	0	0			0	0	1	1	0	0
		1	1	0	0	0	0			1	1	0	0	0	0
		$\mathbf{X}^{(i)}$								$\mathbf{X}^{(j)}$					
		1	1	1	1	1	1			0	0	0	0	1	1
(b)		0	0	1	1	0	0			1	1	1	1	0	0
		1	1	0	0	0	0			1	1	0	0	0	0

Figure 2: Example of a possible one-row crossover move on binary matrices for $K = 3$. (a) Before the crossover: values of \mathbf{X} for chains i and j . (b) After the crossover move (applied to the middle row of the matrices): a potential outcome.

Let $\rho_k^{(i)}$ be the current value of transition probability ρ_k for chain i . By denoting the overall transition probability $p(\mathbf{u}_{t+1}|\mathbf{u}_t)$ for chain i as

$$A^{(i)}(\mathbf{u}_t, \mathbf{u}_{t+1}) := \prod_{k=1}^K \rho_k^{(i)I(u_{k,t} \neq u_{k,t+1})} (1 - \rho_k^{(i)})^{I(u_{k,t} = u_{k,t+1})}$$

we can express

$$\pi_i(\mathbf{Z}_t^{(i)}) = \pi_i(\mathbf{Z}_{t-1}^{(i)}) \cdot \underbrace{\frac{A^{(i)}(\mathbf{v}_{t-1}, \mathbf{v}_t) A^{(i)}(\mathbf{v}_t, \mathbf{u}_{t+1})}{A^{(i)}(\mathbf{u}_{t-1}, \mathbf{u}_t) A^{(i)}(\mathbf{u}_t, \mathbf{v}_{t+1})}}_{c_t^{(i)}} \cdot \frac{p(y_t|\mathbf{v}_t)^{\beta_i}}{p(y_t|\mathbf{u}_t)^{\beta_i}}$$

where $c_t^{(i)}$ denotes the correction term. Using this recursion, we can compute the quantities a_t in terms of a_{t-1} as follows

$$a_t = a_{t-1} \cdot c_t^{(i)} \cdot c_t^{(j)}$$

As the values of a_t can be normalised to sum to one, we can arbitrarily fix the reference value $a_1 \leftarrow 1$. The computation of every correction term is cheap, and the overall complexity for computing all a_t values is $O(KT)$. Moreover, we perform only a one-row crossover, we do not need to consider all K terms in the expression of $A^{(i)}(\mathbf{u}_t, \mathbf{u}_{t+1})$, instead we can focus to the relevant transition probability ρ_k . In this case, the overall complexity reduces even further to $O(T)$.

Furthermore, we typically need to perform the crossover moves only occasionally, so that the ensemble crossover scheme provides a way to improve the poorly mixing samplers for FHMMs at a small extra computational cost. For the rest of the computational time the chains run in parallel. In our experiments, it was sufficient to use two parallel chains: one sampling from the target and the other being tempered.

4 Experiments

In Section 4.1 we demonstrate the proposed method in a toy inference problem that involves sampling from a posterior distribution with two separate modes, while in Section 4.2 we consider a challenging tumor deconvolution example.

4.1 Toy example

To demonstrate the benefit of the ensemble crossover scheme, we consider the following 10-dimensional toy sampling problem, where the target distribution is bimodal.

Let these two modes be $\mathbf{x}_1^0 := (0, 0, \dots, 0, 1, 1, \dots, 1)$ and $\mathbf{x}_2^0 := (1, 1, \dots, 1, 0, 0, \dots, 0)$. We define the probability distribution over all states \mathbf{x} such that $p(\mathbf{x})$ would depend on the Hamming distance from the closest mode, as follows

$$p(\mathbf{x}) \propto 0.1^{\min\{d(\mathbf{x}, \mathbf{x}_1^0), d(\mathbf{x}, \mathbf{x}_2^0)\}}. \quad (2)$$

where $d(\mathbf{x}, \mathbf{y})$ between binary vectors. As illustrated in Figure 3, it is unlikely to observe values \mathbf{x} which are far from both modes.

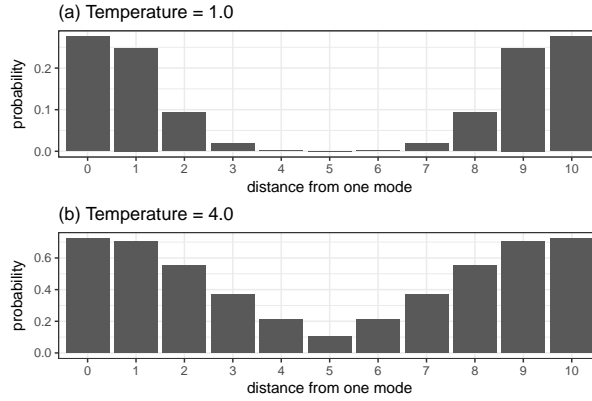


Figure 3: Bimodal probability distribution defined in eq. (2) where $p(\mathbf{x})$ depends on the Hamming distance between \mathbf{x} and the closest mode. We have fixed one mode and show the total (tempered) probability mass (y -axis) for all states for a given Hamming distance (x -axis) from this mode. For the tempered distribution in (b), the density is less peaked and it is more likely to observe values with distance ≈ 5 .

We consider Gibbs sampling on each of the dimensions of \mathbf{x} as a baseline method, and we explore whether our ensemble crossover scheme will improve on this. That is, we run two parallel Gibbs samplers (with temperatures $T_1 = 1.0$ and $T_2 = 4.0$) and carry out a crossover move every 10-th iteration. The target distribution for the tempered chain is illustrated in Figure 3(b). All chains are initialised from the same value (one of the modes) and run for 1000 iterations.

The traces of \mathbf{x} for both methods are shown in Figure 4(ab). Ideally, we would expect a well mixing sampler to sample values around both modes \mathbf{x}_1^0 and \mathbf{x}_2^0 . Standard Gibbs sampler is unable to escape the mode it was initialised from, whereas the ensemble crossover scheme has jumped between the modes on multiple occasions. For each sample, we have also calculated the Hamming distance to the mode from which the samplers was initialised. The Gibbs sampler has not moved across the low probability barrier at Hamming distance ≈ 5 , so the resulting empirical distribution of Hamming distances in Figure 4(c) is very different from the true one in Figure 3(a). In comparison, our ensemble approach enables the sampler to explore both modes. This is achieved via exchanging subsequences with a higher temperature chain, which is sampling from a tempered distribution which is less peaked, as illustrated in Figure 3(b).

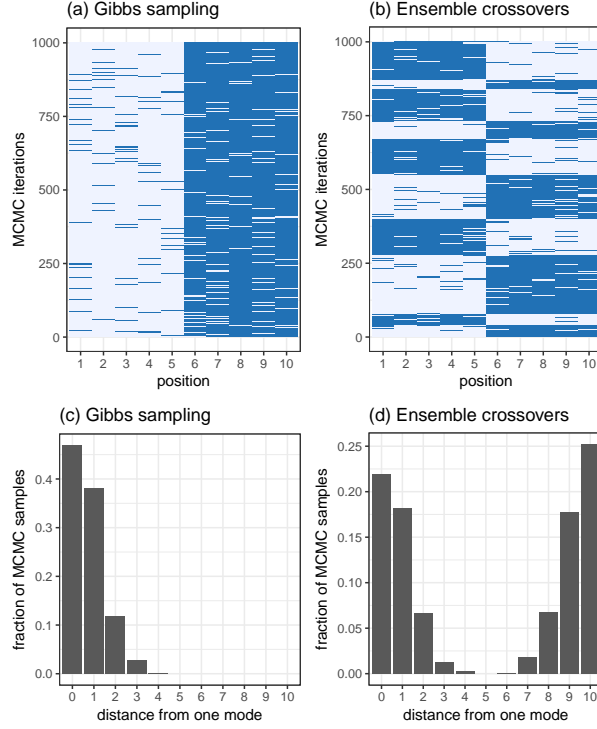


Figure 4: Heatmaps representing the trace plots of \mathbf{x} for (a) Gibbs sampling, (b) ensemble crossover scheme, both targeting the density (2). For each MCMC iteration (y -axis), all 10 elements of the corresponding binary vector \mathbf{x} have been colour coded: light = 0, dark = 1. Based on the traces above, (c) and (d) show the empirical distribution of distances from the mode which both samplers were initialised from. The true target distribution is shown in Figure 3(a).

4.2 Tumor deconvolution example

The following example is motivated by an application in cancer genomics. Certain mutations in the cancer genome result in copy number alterations, i.e. for some genomic regions there can be more or less than two copies of each chromosome. Typically there are various subpopulations present among the cancer cells, and it is of interest to identify the subpopulations to study their phylogeny (Ha et al., 2014; Gao et al., 2016). DNA sequencing technology produces data where read counts have been aggregated over various subpopulations, so an additive FHMM is a natural model to capture these subpopulations.

Lets consider the emission model

$$\mathbf{w} \sim \text{Dirichlet}(\alpha/K, \dots, \alpha/K)$$

$$y_t | \mathbf{x}_t, \mathbf{w}, h \sim \text{Poisson} \left(h \sum_{k=1}^K w_k x_{k,t} \right)$$

where y_t denote the sequence read counts at a locus t and h is the expected sequencing depth. Each w_k corresponds to the fraction of k -th subpopulation ($w_k \geq 0$, $\sum_k w_k = 1$) whose mutation profile is given by the k -th row of \mathbf{X} . Here $x_{k,t} \in \{0, 1\}$ denotes whether the k -th population has a copy number alteration at position k or not.

We note that this is not intended to be a complete model of real-world sequencing data but a device to demonstrate the utility of the proposed ensemble MCMC methods. Further work to construct a sufficiently complex model to capture the variations within real sequencing data, such as single nucleotide polymorphisms, is beyond the scope of this paper and will be developed in future work.

4.2.1 Data generation

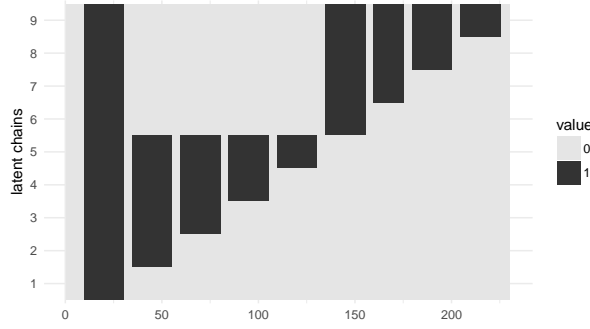


Figure 5: The generated $K \times T$ binary latent matrix \mathbf{X} with $K = 9$ and $T = 230$. Elements $x_{k,t}$ have been colour coded: light = 0, dark = 1.

We generated data according to this Poisson emission model, with $K = 9$ latent chains. The segment lengths were manually specified, resulting in the \mathbf{X} matrix shown in Figure 5. Parameters \mathbf{w} were fixed to $(0.44, 0.04, 0.04, 0.04, 0.04, 0.1, 0.1, 0.1, 0.1)$ and sequencing depth $h = 50$. The \mathbf{w} were chosen such that there would exist multiple underlying matrices \mathbf{X} (i.e. different phylogenetic trees) which could have generated the data, so the resulting posterior will be multimodal.

4.2.2 Comparing various sampling techniques

We carried out inference as follows. To sample from $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$, we represented \mathbf{w} via normalised Gamma random variables, $w_k = \gamma_k / \sum_j \gamma_j$ and then used random walk Metropolis-Hastings on the log-scale of γ_k . To sample from $p(\mathbf{X}|\mathbf{w}, \mathbf{y})$ we compared the following methods:

- "Gibbs" – Gibbs sampling on one row of \mathbf{X} at a time
- "Gibbs + ensemble" – same as previous, but now combined with 1-row crossovers
- "HB" – Hamming Ball sampling with radius $r = 1$
- "HB + ensemble" – same as previous, but now combined with 1-row crossovers

We ran these four samplers for 2,000 iterations, all initialised from the same state. The trace plots of \mathbf{X} are shown in Figure 6 separately for all $K = 9$ latent sequences. As expected, the Gibbs sampler gets stuck and during all 2,000 iterations it moves very little. The Hamming Ball sampler mixes better, but a strong autocorrelation between consecutive states is evident. The ensemble crossover scheme clearly improves mixing of the samplers in both cases, visiting configurations of the state space that were not explored by the standard samplers.

We have chosen $\max_k w_k$ as a summary statistic which represents the current state of the sampler. The ensemble versions of the samplers have visited a wider range of possible $\max_k w_k$ values, as shown in Figure 7. Also the traces of log-likelihood values $\log p(\mathbf{X}|\mathbf{y}_{1:T}, \mathbf{w})$ in Figure 8 indicate that the ensemble crossover scheme has visited areas with a wider range of log-likelihood values, thus moving around the space more freely.

5 Conclusion

We introduced a rejection-free ensemble MCMC method to improve poorly mixing samplers. This is achieved by combining parallel tempering and a novel rejection-free exchange move between pairs of

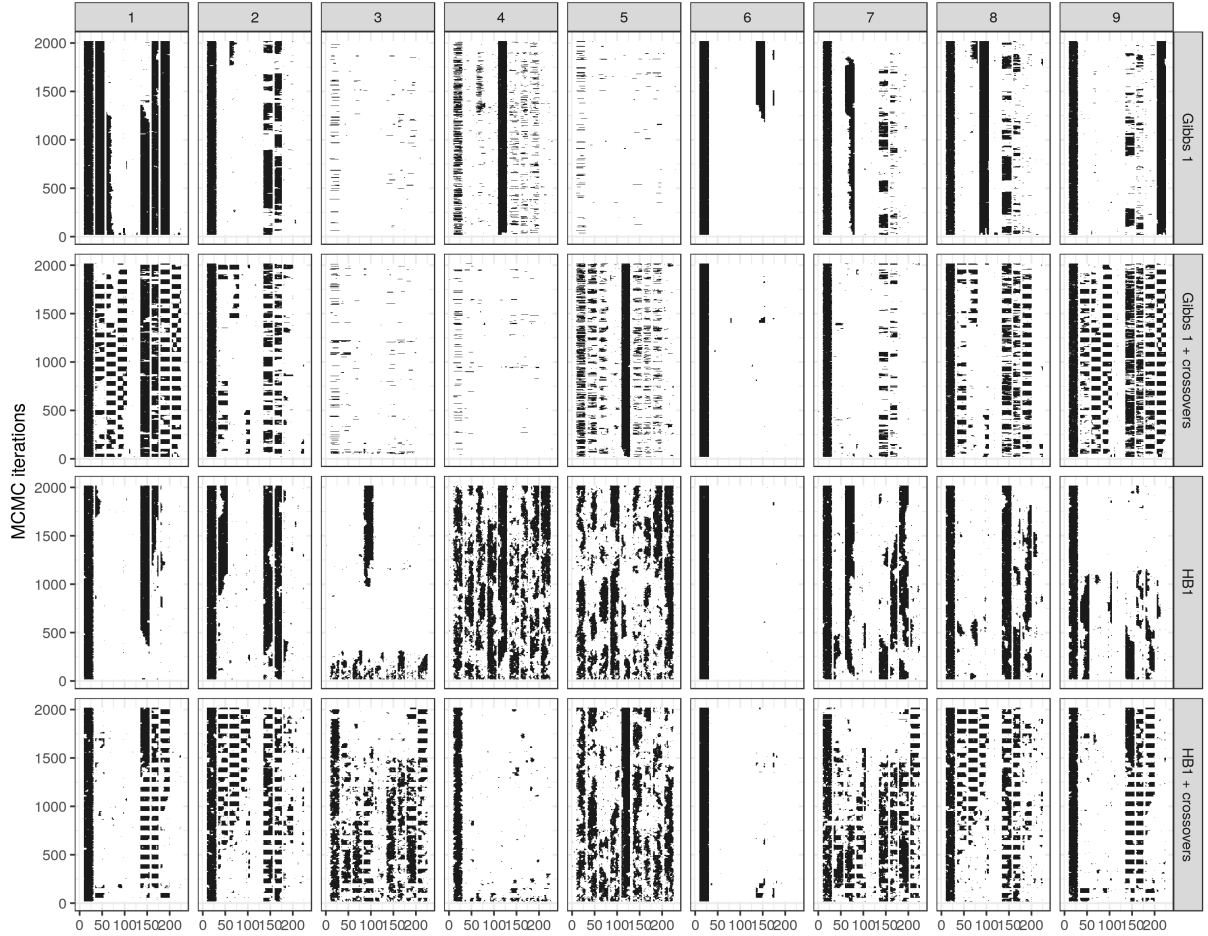


Figure 6: Heatmaps representing trace plots of \mathbf{X} , shown separately for all $K = 9$ latent sequences (in columns) over 2000 MCMC iterations for four sampling methods (shown in rows). For each MCMC iteration (y -axis), we display the elements of k -th latent sequence ($k = 1, \dots, 9$), colour coded light = 0, dark = 1.

chains achieved through an auxiliary variable augmentation. The former allows to have a chain which explores the space freely and does not get stuck, whereas the latter provides an efficient procedure to exchange information between a tempered chain and our target. The rejection-free property combined with the fact that subsequences of varying lengths can be exchanged between the chains, makes our method work even with just two chains in the ensemble scheme. This is typically not feasible with traditional parallel tempering where dense grids of tempered chains are required to achieve sensible acceptance rates.

The proposed method is a general purpose ensemble MCMC approach and here we demonstrated its use on multimodal posterior inference settings in FHMMs. For this model class we described the specifics of how to efficiently implement the crossover move and we demonstrated in a simulation study that the ensemble crossover scheme significantly improves on the efficiency of existing samplers at a low extra computational cost. Our application also illustrates the pragmatic benefits of having a rejection-free scheme in that we do not have to be concerned about parameter tuning in order to achieve optimal acceptance rates in a standard accept-reject framework.

Finally, we expect our technique to be useful in a range of other high-dimensional continuous or

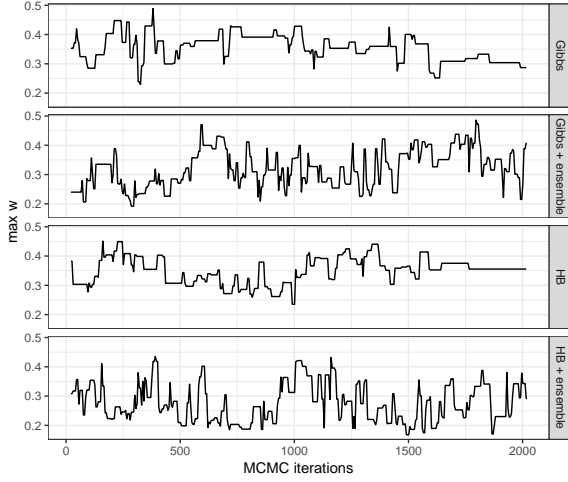


Figure 7: Trace plots of $\max_k w_k$ (y -axis) over 2000 MCMC iterations (x -axis) for four sampling methods (shown in rows).

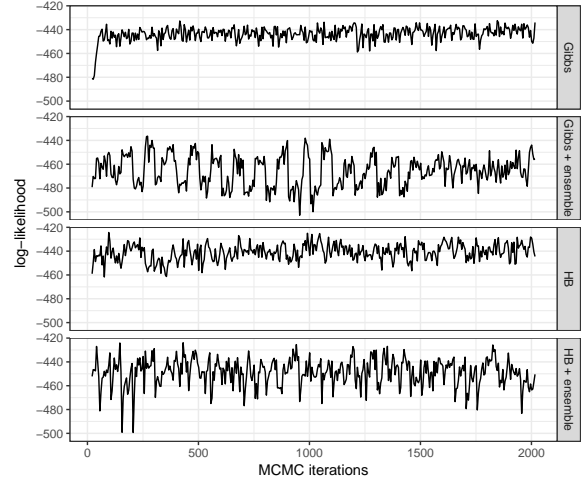


Figure 8: Trace plots of log-likelihood values ($\log p(\mathbf{X}|\mathbf{y}_{1:T}, \mathbf{w})$, y -axis) over 2000 MCMC iterations (x -axis) for four sampling methods (shown in rows).

discrete space sampling problems, such as for spike-and-slab variable selection in regression models and for structural inference in Bayesian neural networks.

Acknowledgements

KM is supported by a UK Engineering and Physical Sciences Research Council Doctoral Studentship. CY was supported by a UK Medical Research Council New Investigator Research Grant (Ref. No. MR/L001411/1) and the Wellcome Trust Core Award Grant Number 090532/Z/09/Z.

References

- Andrieu, Christophe, De Freitas, Nando, Doucet, Arnaud, and Jordan, Michael I. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.
- Crouse, Matthew S, Nowak, Robert D, and Baraniuk, Richard G. Wavelet-based statistical signal processing using hidden markov models. *IEEE Transactions on signal processing*, 46(4):886–902, 1998.
- Earl, David J and Deem, Michael W. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.
- Frellsen, Jes, Winther, Ole, Ghahramani, Zoubin, and Ferkinghoff-Borg, Jesper. Bayesian generalised ensemble markov chain monte carlo. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS), Cadiz, Spain. JMLR: W&CP*, volume 41, 2016.
- Gao, Ruli, Davis, Alexander, McDonald, Thomas O, Sei, Emi, Shi, Xiuqing, Wang, Yong, Tsai, Pei-Ching, Casasent, Anna, Waters, Jill, Zhang, Hong, et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nature Genetics*, 2016.
- Geyer, CJ. *Computing Science and Statistics Proceedings of the 23 Symposium on the Interface; American Statistical Association: New York; p 156*, 1991.
- Ghahramani, Zoubin, Jordan, Michael I, and Smyth, Padhraic. Factorial hidden markov models. *Machine learning*, 29(2-3):245–273, 1997.

- Ha, Gavin, Roth, Andrew, Khattra, Jaswinder, Ho, Julie, Yap, Damian, Prentice, Leah M, Melnyk, Nataliya, McPherson, Andrew, Bashashati, Ali, Laks, Emma, et al. Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome research*, 24(11):1881–1893, 2014.
- Jasra, Ajay, Stephens, David A, and Holmes, Christopher C. On population-based simulation for static inference. *Statistics and Computing*, 17(3):263–279, 2007.
- Liang, Faming and Wong, Wing Hung. Evolutionary monte carlo: Applications to c p model sampling and change point problem. *Statistica sinica*, pp. 317–342, 2000.
- Marchini, Jonathan and Howie, Bryan. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, 2010.
- Neal, Radford M. Mcmc using ensembles of states for problems with fast and slow variables such as gaussian process regression. *arXiv preprint arXiv:1101.0387*, 2011.
- Rabiner, Lawrence and Juang, B. An introduction to hidden markov models. *ieee assp magazine*, 3(1): 4–16, 1986.
- Scott, Steven L. Bayesian methods for hidden markov models. *Journal of the American Statistical Association*, 2002.
- Shestopaloff, Alexander Y and Neal, Radford M. Efficient bayesian inference for stochastic volatility models with ensemble mcmc methods. *arXiv preprint arXiv:1412.3013*, 2014.
- Titsias, Michalis K and Yau, Christopher. Hamming ball auxiliary sampling for factorial hidden markov models. In *Advances in Neural Information Processing Systems*, pp. 2960–2968, 2014.
- Titsias, Michalis K and Yau, Christopher. The hamming ball sampler. *Journal of the American Statistical Association*, (just-accepted), 2016.
- Yau, Christopher. Oncosnp-seq: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics*, 29(19):2482–2484, 2013.